DOCUMENT RESUME

ED 175 929 TM 009 667

AUTHOR Hiatt, Diana Buell; Keesling, J. Ward
TITLE The Dependability of Classroom Observations.
PUB DATE Apr 79
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Classroom Observation Techniques; *Classroom Research; Classroom Techniques; Elementary School Teachers; Informal Assessment; Observation; Primary Education; Research Methodology; *Sampling; Scheduling; *Teacher Behavior; *Test Reliability; *Time
IDENTIFIERS Generalizability Theory; Interrater Reliability

ABSTRACT

A generalizability study of timed observations was conducted in 25 primary grade classes to observe teachers' use of time--for instruction, evaluation of instruction, and classroom management--according to the hour and day observed. Observational methods used by on-site researchers included videotape, checklists, running documentaries, frequency counts, and continuous time allocation by category of behavior. Critical elements of classroom observation were determined to be the teacher, the situation, the occasion, and the evaluator. Teacher behavior in each of ten categories was observed: (1) clerical; (2) preparation of facilities; (3) preparation of materials; (4) planning for instruction; (5) instruction; (6) evaluating instruction; (7) classroom management; (8) supervision; (9) administration; and (10) not teaching. This study appeared to confirm results reported elsewhere: that a total of between four and six observations of teacher behavior, separated over several days, should provide relatively reliable measures of teacher behavior. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

- ED175929

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE DEPENDABILITY OF CLASSROOM OBSERVATIONS

Diana Buell Hiatt

Pepperdine University

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Diana B. Hiatt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

and

J. Ward Keesling

System Development Corporation

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 8-12, 1979.

TM009 667

ABSTRACT

A generalizability study of timed observations in primary classrooms found that different teachers vary greatly in their use of time for instruction, evaluation of instruction, and classroom management according to the hour and day observed. To obtain a generalizable measurement, researchers should select a given time period and record events over several days. Highly trained and selected observers account for an extremely small amount of variance.

Purpose and Issues

Purpose of the Study

Results from major studies evaluating the quality of educational endeavors employing the traditional input-output model reveal little about the mediating variables that affect educational outcomes. There appears to be a growing concern regarding the limitations of such an evaluation or research model and a trend toward seeking information on the factors, or process variables, that affect the various outcome measures (Fisher, 1978; Goodlad, 1979; N.I.E.,1979).

A method of studying process variables that is increasing in popularity is the classroom observation. Observation as a research technique has been employed by clinical psychologists on individuals and by anthropologists on groups for some time. Use of observation in the classroom seems a logical extension of this data-gathering approach. However, classrooms are highly interactive human environments that are complex and demanding to study with the rigor and precision traditionally demanded by educational researchers. Though observations provide a richness of data, such data may include substantial sources of error. Policymakers, who make major decisions and budget allocations based on findings from such observational studies, need to be

aware of the sources of significant variance and the magnitude of such variance.

Errors in the measurement of process variables will attenuate the estimated relationships between process variables and outcomes. Work by Cronbach, et al. (1972) and by Brennan (1977) suggest the need for generalizability studies to assess the effects of various sources of variance on the dependability of classroom observations. Cronbach, et al. argue that human behavior is not reliable or unreliable, but that human behavior can be predicted within certain limits based upon information gleaned from previous observations. They state that the "question of 'reliability' thus resolves into a question of generalization or generalizability (Cronbach, et al., 1972)." A review of recent studies on classroom observation portrays in increasing use of generalizability theory (Borich, et al., 1977). The generalizability study reported here examines the measurement properties of selected process variables which affect educational outcomes.

Since observational research is a costly undertaking, those conducting such research want the most accurate and precise information for the lease amount of expense. The design question to be answered by the generalizability study reported here concerns the allocation of observers to classroom periods. The issue is how best to use observers to obtain precise esti-

mates of the time teachers devote to specific tasks.

Hiatt (1976) conducted an observational study of the effect of aides on the teachers' use of time in primary classrooms. This study provided the data for the generalizability study reported here, yielding evidence about the sources of variability in the observations. The analysis of this data will provide a stronger basis for the design of future observational studies of teacher performance.

Issues Related to Classroom Observational Research

For the purposes of this study, observational method is defined as techniques employed by on-site researchers to record events as they happen. Such techniques include videotape, checklists, running documentaries, frequency counts, or continuous time allocation by category of behavior. Some of the major considerations affecting the measurement of classroom behavior include the following: (1) the clarity or precision in the definition of the categories or items under study; (2) training of the site observers; and (3) adequacy of time sampling both within occasions and across occasions.

One of the first tasks to be accomplished in any study employing the method of classroom observation is to delineate

categories or items under study with clarity and discreteness. Studies of five existing observational schemes revealed problems resulting in ambiguous operational definitions of categories (Borich, et al., 1977). Under such conditions, error in measurement occurs because observers may perceive the categories or items differently. For example, the categories in one observational system included "giving information" and "asking questions." These may seem clear without further definition but observers may discover a gray area such as the use of the rhetorical question, e.g. "Isn't it interesting to note that when c is followed by e, i, or y, it often has the s sound?" Most widely used observational systems include a manual of categories, but Borich, et al. (1977) noted that even with such manuals, observers found gray areas that coders informally delimited or that were eliminated during the training process.

Even under conditions in which observational schemes are unambiguous, there is still the additional requirement that understanding of such schemes be as clear to the persons doing on-site observations. There is no such thing as an "observer-proof" observation instrument. Work by Stallings (1974), Fisher (1978) and Hiatt (1976) attests to the importance of carefully training observers in the conception and perception of the categories under study. Their research shows that adequate training and assurance of high inter-rater agreement prior to the study's

on-site observations result in minimal error created by observers. A subsequent section discusses the procedures Hiatt employed to train observers and to maintain coding standards during data collection.

The classical approach to reliability has been to assign two observers to a classroom and to compute the inter-rater agreement (Rosenshine, 1973). This procedure assumes that what is occurring in the classroom is the true picture of classroom life and that a major source of measurement error is the judges' ability to accurately record such classroom behavior. High inter-rater agreement may assure a clear measurement instrument or coding system but tell nothing about the extent to which the observations recorded represent generalizable characteristics of teachers or students. A highly simplified coding system that assures 100% reliability in recording whether a given behavior did or did not occur on a given occasion provides little information about the frequency with which this behavior occurs in a given classroom across time or across other dimensions, such as subject matter areas.

In developing generalizability theory, Cronbach and his associates, quoting from Medley and Mitzels (1963), conclude that errors in behavior from "one situation or occasion to another far outweigh error arising from failure of two obser-

vers to agree exactly in their records of the same behavior" (Cronbach, et al., 1972:190). They caution that inter-rater agreement is only a small part of the reliability problem.

The present study employs a broader conception of reliability that will provide an assessment of the generalizability of differences among recorded classroom behaviors rather than simply the differences among observers' perceptions. A model proposed by McGaw, et al. (1972) suggests that the critical elements of classroom observation can be identified as the teacher, the situation, the occasion and the judge (or rater). This model suggests that the use of judges over a number of occasions may prove more reliable in recording differences across teachers than a large number of judges observing one occasion. Clearly, then, the sampling of occasions and situations is of great importance. The following two examples illustrate the importance of including occasions and situations as sources of variance in assessing the dependability of observations.

Erlich and Borich (1976) applied generalizability theory to the analysis of teacher questioning behavior in reading. Their analysis revealed that three to five occasions of observation were required to reach a .7 generalizability coefficient. The number of occasions varied by type of question.

Rowley (1975) reanalyzed classroom observational data by breaking down 50 minute observational periods into 10 minute segments. His findings showed that there were higher generalizability coefficients resulting from shorter periods of observation over more situations. If a person wants to obtain a complete picture of classroom behavior, he needs to observe over several periods of time and repeat the situation. His data suggested that four to six observations, each 30 minutes long, would provide a fairly stable picture of a given teacher, classroom or set of pupils. Then the fluctuations noted between classrooms would be true variance among teachers generalized across occasions and situations.

The study reported here attempted to clearly delineate observation categories and provide sufficient training of observers prior to on-site data collection. This generalizability study will examine selected sources of variance in observations and attempt to provide information regarding allocation of observers to classroom periods.

Design of the Study

Study Sample

Twenty-five classrooms were randomly selected from five

different schools in a large western city. The selection was designed to achieve homogeneity among the classrooms. Each classroom was graded and self-contained. An analysis of the classrooms showed nearly equal division among the three grade levels of the selected populations and and equal sampling from various socioeconomic levels, excluding very high SES and very low SES. The sample included seven first grade classes, nine second grade classes, and nine third grade classes. An analysis of the teachers' years of educational training and experience in the classroom reveals a normal distribution for both variables.

Description of Design

A study by Hiatt (1976) focusing on the effects of teacher aides on teaching behavior provided the raw data for this study. The data met certain assumptions in generalizability theory, namely:

- The universe is unambiguously described so appropriate categories and situations cna be clearly identified.
- Situations are independent so that a score in one situation does not depend on a previous score or observation in another situation.
- Data is on an interval scale.

This study employed a four facet mixed and partially nested design.

The selected facets affecting variance in the classroom observations were:

Individual teacher	t	Randomly selected
Day observed	d	Randomly selected
Hour of the day	h	Fixed
Individual Observer	o	Randomly selected

Time in minutes and seconds was chosen as the unit of measurement. The observations occurred naturalistically within the classrooms for a continuous 180 minutes beginning with the formal opening of each school day. Two observations were scheduled for each classroom. Teachers were asked to eliminate from the randomly selected days any day that would not provide "typical" classroom activities, and alternate days were assigned. Observations occurred across all days of the week, and no class was visited twice on the same day of the week. There were several days between the two observation days.

Categories of Teaching Activities

Various observational schemes have been developed and studied which analyze certain aspects of the act of teaching. Borich, et al. (1977) have analyzed the generalizability of such process measures. None of these encompass the total teaching behavior which may occur within a classroom. The need for a conceptual framework within which the full-spectrum of teaching behaviors

could be observed became apparent. Figure 1 presents a model developed by Hiatt (1976) to describe the activities of ten inclusive categories of teaching: clerical, preparation of facilities, preparation of materials, planning for instruction, instruction, evaluating instruction, classroom management, supervision, administration, and out (not teaching).

The following definitions describe the nature or essence of each category of teaching behavior:

1. Clerical work is the performance of recording and accounting functions.
2. Preparation of facilities is the organization and preparing of furniture and equipment within the classroom for instruction.
3. Preparation of materials is the preparation of instructional materials or the assemblage of materials from outside sources.
4. Planning for instruction is the organizing of the content of instruction.
5. Instruction is the activity of imparting skills and knowledge to learners.
6. Evaluating instruction is the assessment of instruction that has taken place.
7. Classroom management is the organization, control, and care of a classroom of children.

8. Supervision is the overseeing and monitoring of the behavior of others.
9. Administration is the performance of activities that deal with the assignment and control of the work of other adults, the operation of the affairs of the school outside of the classroom, and the participation in district, state and federal programs.

Explanation of the theoretical and empirical basis for these categories can be found in the complete report of this study (Hiatt, 1976).

Training of Observers

A team of thirteen graduate students in the field of education at two branches of a large western university were trained to observe in classrooms using a stop-watch and the categories of the model of teaching presented above. These students participated in a series of four highly structured two-hour training sessions interspersed with naturalistic observations in classrooms.

During the first two-hour training session, the students were instructed in the use of the stop-watch and in the Hiatt model of teaching. An examination was given on accurate use of the stop-watch and an understanding of the categories of teaching. A short videotape of a classroom was shown and the students

attempted to record the teacher's use of time by category. Following the first training session, the students observed in classrooms as part of their regular teacher training program. They recorded the classroom teachers' use of time using the stop-watch and the ten categories for thirty minutes.

At the beginning of the second session, sources of confusion in allocating time were discussed and a review of the ten categories took place. The students noted that teachers tended to follow two basic teaching patterns. The first is commonly termed the diagnostic-prescriptive model in which teachers evaluate the level at which the student is operating (diagnose), instruct the student(s) in the desired behavior, and then plan future instructional activities with the student (prescribe). The second pattern was the traditional procedure in which the teacher instructed (gave a lesson), planned future student instructional activities (usually assignment of seatwork), and evaluated instruction toward the end of the instructional period. The activities related to teaching, especially classroom management, may intersperse that basic instructional pattern or occur separately (as during playground supervision, collecting lunch money, etc.).

A thirty minute videotape was shown to the students at this second session. The students independently recorded the teaching time by category. The data from each student was compared with a master sheet prepared by the researchers. The same videotape

was shown again, and there was a significant gain in student reliability. The students were assigned to record two thirty-minute periods during their observation visits to classrooms.

The third and fourth sessions were comprised of viewing additional videotapes until the student recording behavior demonstrated an agreement with the master sheet for each videotape of over 90 percent. After the third session the students returned to record one full hour of teaching by category in their assigned classrooms.

Following the fourth training session the students were assigned to observe in the target classrooms of the study in randomly assigned pairs. For this study, identical Breitling stop-watches were used by all observers. This stop-watch had a 60-second sweep and an ability to record consecutively for 30 minutes. The stop-watch could be silently returned to the start position so that teachers and pupils would not be disturbed by a clicking noise. Records showed that teachers change categories often, and observers recorded from fifty to one hundred changes within an hour.

The observers were requested to seat themselves in opposite sides of each classroom and remain as unobtrusive as possible. They were to remain seated unless movement was necessary to observe activities of the teacher. Each observer recorded independently of the other.

For 180 consecutive minutes the observers recorded teaching behavior by category using special recording sheets which had the categories of teaching behavior at the top of ten columns with 25 boxes under each category to record discrete amounts of time. The observers began a new recording sheet at the start of each hour. Each recording sheet was precoded to insure confidentiality and correctness of the data. At the end of the 180 minutes of recording, each observer tallied the amount of time by category and then summed across the columns to make sure 60 minutes were recorded for each hour. The primary researcher was present at each school on the days of observation and made periodic checks on the accuracy of the observers' recording. Table 1 presents the average use of classroom time for each hour of observation.

Insert Table 1 about here

Data Analysis

Because the fixed effects of aides and grade levels which were part of the original study are not of interest in the generalizability study, the data used in the analyses reported here are the residuals from fitting a two-way analysis of variance model which removed the effects of the aide (vs no aide)

and grade-level factors, and their interaction. The residuals were analyzed using a components of variance model in which the teacher was one factor, the hour of day was a factor crossing teachers and the day of observation was nested within the teacher-by-hour interaction.

This model is similar to the teacher-by-situation model of McGaw, Wardrop and Bunda (1972). However, the pairs of observers were allocated to teacher and day at random, so they could not cross these factors in the way proposed by McGaw, **et al.** Furthermore, the status of the hour factor is not identical to their situation factor. On the one hand, it can be seen as random, (as was the McGaw, et al., Situation factor), representing a source of variation in the observations. This would be of interest if only one randomly chosen hour was to be observed. On the other hand, the three morning hours constitute an exhaustive classification of the time to be observed and could be considered as fixed, as they represent the total set of hours over which generalizations are to be made. Furthermore, if all three hours are to be observed and inferences are to be made about the entire teaching performance the hour factor may be eliminated. A conservative position has been adopted in which hours are treated as a random factor. The tables presented should prove to be useful both to the researcher who wants to sample hours and the researcher who will observe for the entire morning.

The following univariate model was developed to explain the variation in the classroom observations:

$$(1)\quad V[y] = V[t] + V[h] + V[th] + V[d(th)] + V[E]$$

where

$V[y]$ is the variance observed in number of minutes

$V[t]$ is the variance due to teachers

$V[h]$ is the variance due to hour

$V[th]$ is the variance due to teacher-by-hour interaction

$V[d(th)]$ is the variance due to day or occasion, within the teacher-by-hour interaction

$V[E]$ is the variance among the observers and observer interactions with teachers and hours

Generalizability coefficients were estimated by ratios of estimated variance components:

$$(2)\quad \hat{\rho}_t^2 = \hat{V}[t] / (\hat{V}[t] + \hat{V}[d(th)]/d + \hat{V}[E])$$

The effects on these coefficients of increasing the number of days observed and/or the number of observers will be presented. Four of the ten activities which were observed in the present study -- Instruction Time, Evaluation Time, Management Time, and Out Time -- were selected for this generalizability study.

Out Time was chosen because it was felt that nonteaching

time would be very accurately assessed by all observers. Thus, it could form a point of reference to use in assessing the generalizability of the other observations. In addition, there are clear, large hour effects on Out Time (due to recess scheduled in the second hour). These effects enable one to make informative comparisons among the variance components for hour across the four variables.

Insert Table 2 about here

The data in Table 2 clearly show that the variation among hours and that among days (within teacher-by-hour) are very important determiners of the over-all variance in the observations. The error variance (contributed by observer differences) is very small, by comparison. It can be concluded that none of these variables requires more than one observer per observation period; the gain in precision would be miniscule, especially compared to the cost of employing two observers for each observation.

The coefficient of prime interest is the generalizability of the measurement of teacher differences. The appropriate coefficient may be estimated by forming the ratio given in (2) which is the reliability of the observations of teacher behaviors within any particular hour. Table 3 shows the estimated reliability coefficients for varying numbers of days of observation ($\underline{d}$, in formula 2).

Insert Table 3 about here

There are virtually no measurable differences in Out Time between teachers within hours. Evaluation time also does not vary greatly among teachers within hours. Consequently, it is very difficult to reliably measure teacher differences on these two variables. By making six observations of instruction and classroom management, it is possible to assess teacher differences (within hour) fairly reliably.

The result for Out Time is interesting because of the a priori belief in the accuracy of its measurement. Because teachers show no variation on this variable, teacher differences cannot be assessed. However, differences across hours in the amount of Out Time can be very reliably detected: with one day of observation the reliability coefficient for hours (analogous to that for teachers, substituting $V[h]$ for $V[t]$ in formula 2) is .72.

Table 4 presents the variance components when the three morning hours are considered as a unit of observation. In this case, the components associated with hours are eliminated.

Insert Table 4 about here

Table 5 gives the estimated generalizability coefficients for the various numbers of days of observation.

Insert Table 5 about here

If the entire morning of teaching is to be observed as one unit, then acceptable generalizability coefficients may be obtained with four to six days of observation, for all variables except Instruction Time. The teachers in this study apparently did not differ much in their total teaching time over a morning, while they did differ in their total morning Out Time.

Implications for the Design of Experiments

The estimated variance components and generalizability coefficients reported above have interesting implications for the design of experiments using teachers as the units of analysis. In a simple experiment designed to contrast an experimental treatment with a control, teachers would be randomly assigned to these two conditions. The pooled, within-group variability of observed scores for the teachers would determine the precision of the experiment, that is, the ability of the experiment to detect differences between the treatment conditions. If one wanted to examine the effects of an experimental treatment on

the total amount of time teachers spend out of class during the three morning hours, it seems clear that teacher-to-teacher differences will be the most important source of variance in the observations. To improve the precision of this hypothetical experiment one would need to sample additional teachers in each treatment condition.

If a similar experiment were designed to assess the effects on instructional time, it is clear that the major source of variation in observations is the day-to-day variation within teacher. Here, one could enlist fewer teachers in the experiment, but observe each one more frequently. While the additional observations might not yield measures precise enough to enable one to reliably distinguish among teachers within either treatment condition, they would increase the precision of the experiment by reducing the variability of the observations in each group.

It is important to note that when observations are part of an experimental design, the precision of measurement of a difference between group means is a function of the reliability or generalizability of the observation for each experimental unit as well as the number of experimental units in each group. Even if the observation for each unit had very low generalizability, as is the case for Instructional Time, group differences may be detected by increasing the number of units assigned to the treatment conditions. The designer of an experiment thus faces a

cost-effectiveness trade-off: whether to observe more classrooms with lesser reliability, or fewer classrooms with greater reliability. Generally speaking, there are practical limits to the number of experimental units that the experimenter can deal with. These limits are imposed by the work that must be done to obtain the cooperation of even a modest number of teachers and to train them in the experimental conditions to be employed. The exact nature of the trade-off is determined by the desired precision of estimation of the experimental effect.

Summary and Conclusions

A generalizability study of observations of teachers' use of time in the classroom was conducted using a random sample of 25 teachers. Each teacher was observed by two observers on two different days for the first three hours of the school day. The generalizability study undertaken indicated that trained observers recorded highly consistent observations of the amounts of time teachers devoted to well-defined categories of teaching activities. The variations across days and across teachers depend upon the characteristic of teacher behavior being observed. Teacher-to-teacher variation in total out-of-classroom time is very large compared to the day-to-day variation in this behavior. These relative magnitudes are different for the measure of time spent in instruction. Because of this difference in relative magnitudes, it seems reasonable to conclude that researchers should conduct generalizability studies involving their choice

of variables before they embark on decision studies.

In general, this study seems to confirm the findings of other researchers that somewhere between four and six observations of a teacher's behavior, separated over several days, should provide relatively reliable measures of teacher behavior. These observations can be conducted by a single observer when there is a delineation of categories or items under study and the training of the observers assures their consistent use of the observational schedule.

When the researcher plans to do an experimental study, the precision of the experiment may be increased by increasing the number of experimental units employed in the design (within practical limits), or by increasing the precision of the observations of each one of a smaller number of units. The components of variance estimated as part of the computation of the generalizability coefficients will be useful in assessing which design optomizes the use of the researcher's resources.

Table 1

Teacher Use of Classroom Time

in Mean Number of Minutes

Category of Teaching Activity	Hour 1	Hour 2	Hour 3
Nonteaching	.16	12.80	1.33
Clerical	3.69	.42	.87
Preparation of facilities	.61	.71	1.08
Preparation of materials	3.44	2.64	2.80
Planning for instruction	7.95	4.91	6.76
Instruction	14.06	9.66	12.78
Evaluation of instruction	14.55	9.20	16.65
Classroom management	9.69	10.49	12.98
Supervision	3.97	7.56	3.27
Administration	1.89	1.61	1.47

Table 2

Estimated Components of Variance in The Model for Observation of Teacher Use of Time

Dependent variable	V[t]	V[h]	V[th]	V[d (th)]	V[E]
Instruction	12.4	4.2	4.5	35.0	0.27
Evaluation	2.4	13.6	15.1	30.0	0.56
Management	6.9	2.6	2.7	13.3	0.34
Nonteaching	0.0	48.0	10.1	18.5	0.00

Table 3

Reliability of Classroom Observations of Teacher Behaviors Within Hour

Dependent variable	Number of Days Observed				
	1	2	4	6	8
Instruction	.26	.41	.58	.67	.73
Evaluation	.07	.14	.23	.30	.36
Management	.34	.50	.65	.73	.78
Nonteaching	.00	.00	.00	.00	.00

Table 4

Estimate of Variance Components in Teacher Use of Total Time for Three Consecutive Morning Hours

Dependent variable	Estimated Variance Components		
	V t	V d(t)	V E
Instruction	0.04	4.90	0.06
Evaluation	31.50	36.50	4.80
Management	148.10	116.50	14.80
Nonteaching	6.36	0.00	0.04

Table 5

Coefficients of Generalizability of Teacher Use of Time

Dependent variable	Number of Days Observed				
	1	2	4	6	8
Instruction	.01	.02	.03	.04	.06
Evaluation	.43	.58	.69	.74	.77
Management	.53	.67	.77	.81	.83
Nonteaching	.99	.99	.99	.99	.99

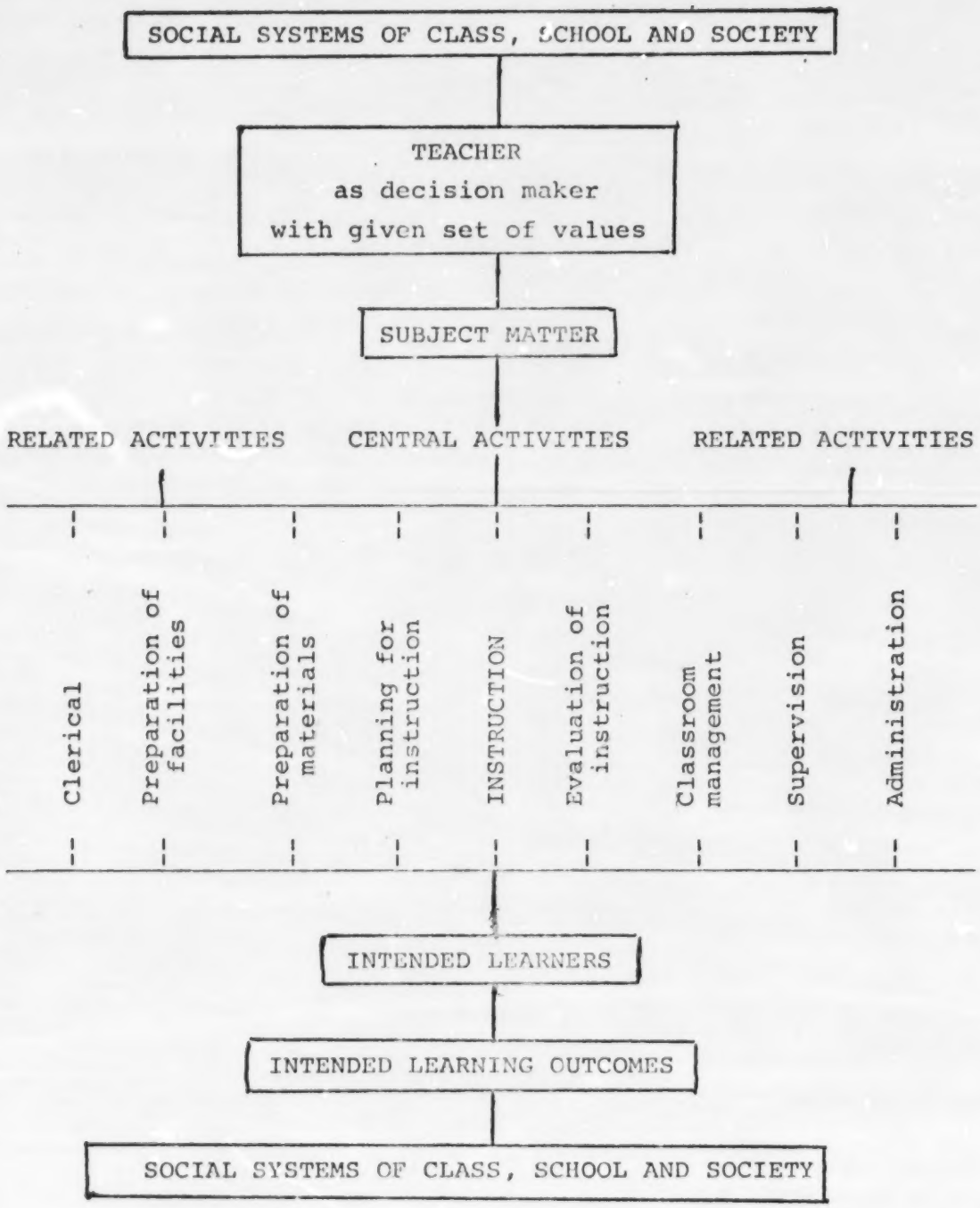

Figure 1. Categories of central and related activities of teaching.

REFERENCES

Borich, G. and others. Classroom observation data: Is it valid? Is it generalizable? A compendium of papers. Austin, Texas: Texas University 1977, (ERIC Document Reproduction Service No. ED 137 371).

Brennan, R. L. Generalizability Analyses: Principles and Procedures. Iowa City, Iowa: American College Testing Program, 1977, (ERIC Document Reproduction Service No. ED 146 215).

Brown, B. B., Mendenhall, W. and Beaver, R. The reliability of observations of teachers' classroom behavior. Journal of experimental education, 1968, 36, 1-10.

Cronbach, L. J. and others. The dependability of behavioral measurements: A theory of generalizability for scores and profiles. New York: John Wiley and Sons, Inc., 1972.

Erlich, O. and Borich, G. Generalizability of teachers process behaviors during reading instruction. Austin, Texas: Texas University, 1976. (ERIC Document Reproduction Service No. ED 142 586)

Fisher, C. W. et al. Teaching and learning in the elementary school: a summary of the beginning teacher evaluation study. Beginning teacher evaluation study, report VII. San Francisco: Far West Laboratory, 1978.

Goodlad, J. et al. A study of schooling: research in progress, Symposium at the Annual Meeting of the American Educational Research Association, San Francisco. 1979.

Hiatt, D. B. The effect of teacher aides on use of time and individualization of instruction by primary teachers: Dissertation abstracts,December, 1976.

McGaw, B.,Wardrop, J. L., & Bunda, M. A. Classroom observation schemes: where are the errors? American Educational Research Journal, 1972, 9, 13-27.

Medlay, D. M. and Mitzel, H. Application of analysis of variance to estimation of the reliability of observations of teachers' classroom behavior. Journal of experimental education, 1958, 27, 23-35.

National Institute of Education, Teaching and Research Grants Announcement, March 1979.

Rawley, G. A rationale for assessing the reliability of an observational measure. Paper presented at the Annual Meeting of the American Educational Research Association, March 30 - April 3, 1975. (ERIC Document Reproduction Service No. ED 104 937)

Rosenshine, B. and Furst, N. G. The use of direct observation to study teaching. In P. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand MacNally, 1973, 122-183.

Shavelson, R. and Dempsey, N. Generalizability of measures of teaching behavior. Review of educational research, 1976, 46, 553-611.

Stallings, J. A. and Kaskowitz, D. H. Follow-through classroom observation, 1972-1973. Menlo Park, California: Stanford Research Institute, 1974.